Specification

Text-Processing Method, Program, Program Recording Medium, and Device Thereof

Technical Field

The present invention relates to a text-processing method of segmenting a text document comprising character strings or word strings for each semantic unit, i.e., each topic, a program, a program recording medium, and a device thereof.

Background Art

A text-processing method of this type, a program, a program recording medium, and a device thereof are used to process enormous and many text documents so as allow a user to easily obtain desired information therefrom by, for example, segmenting and classifying the text documents for each semantic content, i.e., each topic. In this case, a text document is, for example, a string of arbitrary characters or words recorded on a recording medium such as a magnetic disk. Alternatively, a text document is the result obtained by reading a character string printed on a paper sheet or handwritten on a tablet by using an optical character reader (OCR), the result obtained by causing a speech recognition device to recognize speech waveform signals generated by utterances of persons, or the like. In general, most of

signal sequences generated in chronological order, e.g., records of daily weather, sales records of merchandise in a store and records of commands issued when a computer is operated, fall within the category of text documents.

Conventional techniques associated with this type of text-processing method, program, program recording medium, and device thereof are roughly classified into two types of techniques. These two types of conventional techniques will be described in detail with reference to the accompanying drawings.

According to the first conventional technique, an input text is prepared as a word sequence o_1, o_2,..., o_T, and statistics associated with word occurrence tendencies in each section in the sequence are calculated. A position where an abrupt change in statistics is seen is then detected as a point of change in topic. For example, as shown in FIG. 5, a window having a predetermined width is set for each portion of an input text, the occurrence counts of words in each window are counted, and the occurrence frequencies of the words are calculated in the form of a polynomial distribution. If a difference between two adjacent windows (windows 1 and 2 in FIG. 5) is larger than a predetermined threshold, it is determined that a change in topic has occurred at the boundary of the two windows. As a difference between two windows, for

example, the KL divergence between the polynomial distributions calculated for the respective windows can be used as represented by, for example, expression (1):

$$\sum_{i=1}^{L} a_i \log \frac{a_i}{b_i} \qquad \ldots(1)$$

where a_i and b_i ($i = 1, \ldots, L$) are polynomial distributions representing the occurrence frequencies of words corresponding to windows 1 and 2, respectively, $a_1 + a_2 + \cdots + a_L = 1$ and $b_1 + b_2 + \cdots + b_L = 1$ hold, and L is the vocabulary size of the input text.

In the above operation, a so-called unigram is used, in which statistics in each window are calculated from the occurrence frequency of each word. However, the occurrence frequency of a concatenation of two or three adjacent words or a concatenation of an arbitrary number of words (a bigram, trigram, or n-gram) may be used. Alternatively, each word in an input text may be replaced with a real vector, and a point of change in topic can be detected in accordance with the moving amount of such a vector in consideration of the co-occurrence of non-adjacent words (i.e., simultaneous occurrence of a plurality of non-adjacent words in the same window), as disclosed in Katsuji Bessho, "Text Segmentation Using Word Conceptual Vectors", Transactions of Information Processing Society of Japan, November 2001, Vol. 42, No. 11, pp. 2650 - 2662 (reference 1).

According to the second conventional technique, statistical models associated with various topics are prepared in advance, and an optimal matching between the models and an input word string is calculated, thereby obtaining a topic transition. An example of the second conventional technique is disclosed in Amaral et al., "Topic Detection in Read Documents", Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries, 2000 (reference 2). As shown in FIG. 6, in this example of the second conventional technique, statistical models for topics, e.g., "politics", "sports", and "economy", i.e., topic models, are formed and prepared in advance. A topic model is a word occurrence frequency (unigram, bigram, or the like) obtained from text documents acquired in large amounts for each topic. If topic models are prepared in this manner and the probabilities of occurrence of transition (transition probabilities) between the topics are properly determined in advance, a topic model sequence which best matches an input word sequence can be mechanically calculated. As easily understood by replacing an input word sequence with an input speech waveform and replacing a topic model with a phoneme model, a topic transition sequence can be calculated in the manner of DP matching by using a calculation method such as frame-synchronized beam search as in many conventional techniques associated

with speech recognition.

According to the above example of the second conventional technique, statistical topic models are formed upon setting topics which can be easily understood by intuition, e.g., "politics", "sports", and "economy". However, as disclosed in Yamron et al., "Hidden Markov Model Approach to Text Segmentation and Event Tracking", Proceedings of International Conference on Acoustic, Speech and Signal Processing 98, Vol. 1, pp. 333 - 336, 1998 (reference 3), there is also a technique of forming topic models irrelevant to human intuition by applying some kind of automatic clustering technique to text documents. In this case, since there is no need to classify in advance a large amount of text documents for each topic to form topic models, the labor required is slightly smaller than that in the above technique. This technique is however the same as that described above in that a large-scale text document set is prepared, and topic models are formed from the set.

Disclosure of Invention

Problem to be Solved by the Invention

Both the above first and second conventional techniques have a few problems.

In the first conventional technique, it is difficult to optimally adjust parameters such as a threshold associated with a difference between windows and a window width which defines a count range of word

occurrence counts. In some case, a parameter value can be adjusted for desired segmentation of a given text document. For this purpose, however, time-consuming operation is required to adjust a parameter value in a trial-and-error manner. In addition, even if desired operation can be realized with respect to a given text document, it often occurs that expected operation cannot be realized when the same parameter value is applied to a different text document. For example, as a parameter like a window width is increased, the word occurrence frequencies in the window can be accurately estimated, and hence segmentation processing of a text can be accurately executed. If, however, the window width is larger than the length of a topic in the input text, the original purpose of performing topic segmentation cannot be obviously attained. That is, the optimal value of a window width varies depending on the characteristics of input texts. This also applies to a threshold associated with a difference between windows. That is, the optimal value of a threshold generally changes depending on input texts. This means that expected operation cannot be implemented depending on the characteristics of an input text document. Therefore, a serious problem arises in actual application.

In the second conventional technique, a large-scale text corpus must be prepared in advance to form topic models. In addition, it is essential that

the text corpus has been segmented for each topic, and it is often required that labels (e.g., "politics", "sports", and "economy") have been attached to the respective topics. Obviously, it takes much time and cost to prepare such a text corpus in advance. Furthermore, in the second conventional technique, it is necessary that the text corpus used to form topic models contain the same topics as those in an input text. That is, the domains (fields) of the text corpus need to match those of the input text. In the case of this conventional technique, therefore, if the domains of an input text are unknown or domains can frequently change, it is difficult to obtain a desired text segmentation result.

It is an object of the present invention to segment a text document for each topic at a lower cost and in a shorter time than in the prior art.

It is another object to segment a text document for each topic in accordance with the characteristics of the document independently of the domains of the document.

Means of Solution to the Problem

In order to achieve the above objects, a text-processing method of the present invention is characterized by comprising the steps of generating a probability model in which information indicating which word of a text document belongs to which topic is made

to correspond to a latent variable and each word of the text document is made to correspond to an observable variable, outputting an initial value of a model parameter which defines the generated probability model, estimating a model parameter corresponding to a text document as a processing target on the basis of the output initial value of the model parameter and the text document, and segmenting the text document as the processing target for each topic on the basis of the estimated model parameter.

In addition, a text-processing device of the present invention is characterized by comprising temporary model generating means for generating a probability model in which information indicating which word of a text document belongs to which topic is made to correspond to a latent variable and each word of the text document is made to correspond to an observable variable, model parameter initializing means for outputting an initial value of a model parameter which defines the probability model generated by the temporary model generating means, model parameter estimating means for estimating a model parameter corresponding to a text document as a processing target on the basis of the initial value of the model parameter output from the model parameter initializing means and the text document, and text segmentation result output means for segmenting the text document as the processing target

for each topic on the basis of the model parameter estimated by the model parameter estimating means.

Effects of the Invention

According to the present invention, it does not take much trouble to adjust parameters in accordance with the characteristics of a text document as a processing target, and it is not necessary to prepare a large-scale text corpus in advance by spending much time and cost. In addition, the present invention can accurately segment a text document as a processing target for each topic independently of the contents of the text document, i.e., the domains.

Brief Description of Drawings

Fig. 1 is a block diagram showing the arrangement of a text-processing device according to an embodiment of the present invention;

Fig. 2 is a flowchart for explaining the operation of the text-processing device according to an embodiment of the present invention;

Fig. 3 is a conceptual view for explaining a hidden Markov model;

Fig. 4 is a block diagram showing the arrangement of a text-processing device according to another embodiment of the present invention;

Fig. 5 is a conceptual view for explaining the first conventional technique; and

Fig. 6 is a conceptual view for explaining the

second conventional technique.

Best Mode for Carrying Out the Invention

First Embodiment

The first embodiment of the present invention will be described next in detail with reference to the accompanying drawings.

As shown in Fig. 1, a text-processing device according to this embodiment comprises a text input unit 101 which inputs a text document, a text storage unit 102 which stores the input text document, a temporary model generating unit 103 which generates one or a plurality of models each describing the transition between topics (semantic units) of the text document and in which information indicating which word of the text document belongs to which topic is made to correspond to a latent variable (a variable which cannot be observed) and each word of the text document is made to correspond to an observable variable (a variable which can be observed), a model parameter initializing unit 104 which initializes the value of each model parameter which defines each model generated by the temporary model generating unit 103, a model parameter estimating unit 105 which estimates the model parameter of the model initialized by the model parameter initializing unit 104 by using the model and the text document stored in the text storage unit 102, an estimation result storage unit 106 which stores the parameter estimation result

obtained by the model parameter estimating unit 105, a model selecting unit 107 which selects a parameter estimation result on one model from parameter estimation results on a plurality of models if they are stored in the estimation result storage unit 106, and a text segmentation result output unit 108 which segments the input text document in accordance with the parameter estimation result on the model selected by the model selecting unit 107 and outputs the segmentation result. Each unit can be implemented by being operated by a program stored in a computer or by reading the program recorded on a recording medium.

In this case, as described above, a text document is a string of arbitrary characters or words recorded on a recording medium such as a magnetic disk. Alternatively, a text document is the result obtained by reading a character string printed on a paper sheet or handwritten on a tablet by using an optical character reader (OCR), the result obtained by causing a speech recognition device to recognize speech waveform signals generated by utterances of persons, or the like. In general, most of signal sequences generated in chronological order, e.g., records of daily weather, sales records of merchandise in a store and records of commands issued when a computer is operated, fall within the category of text documents.

The operation of the text-processing device

according to this embodiment will be described in detail next with reference to Fig. 2.

The text document input from the text input unit 101 is stored in the text storage unit 102 (step 201). Assume that in this case, a text document is a word sequence which is a string of T words, and is represented by o_1, o_2,..., o_T. A Japanese text document, which has no space between words, may be segmented into words by applying a known morphological analysis method to the text document. Alternatively, this word string may be formed into a word string including only important words such as nouns and verbs by removing postpositional words, auxiliary verbs, and the like which are not directly associated with the topics of the text document from the word string in advance. This operation may be realized by obtaining the part of speech of each word using a known morphological analysis method and extracting nouns, verbs, adjectives, and the like as important words. In addition, if the input text document is a speech recognition result obtained by performing speech recognition of a speech signal, and the speech signal includes a silent (speech pause) section, a word like <pause> may be contained at the corresponding position of the text document. Likewise, if the input text document is a character recognition result obtained by reading a paper document with an OCR, a word like <line feed> may be contained at a

corresponding position in the text document.

Note that in place of a word sequence (unigram) in a general sense, a concatenation of two adjacent words (bigram), a concatenation of three adjacent words (trigram), or a general concatenation of n adjacent words (n-gram) may be regarded as a kind of word, and a sequence of such words may be stored in the text storage unit 102. For example, the storage form of a word string comprising concatenations of two words is expressed as (o_1, o_2), (o_2, o_3),..., (o_{T-1}, o_T), and the length of the sequence is represented by T-1.

The temporary model generating unit 103 generates one or a plurality of probability models which are estimated to generate an input text document. In this case, a probability model or model is generally called a graphical model, and indicates models in general which are expressed by a plurality of nodes and arcs which connect them. Graphical models include Markov models, neural networks, Baysian networks, and the like. In this embodiment, nodes correspond to topics contained in a text. In addition, words as constituent elements of a text document correspond to observable variables which are generated from a model and observed.

Assume that in this embodiment, a model to be used is a hidden Markov model or HMM, its structure is a one-way type (left-to-right type), and an output is a

sequence of words (discrete values) contained in the above input word string. According to a left-to-right type HMM, a model structure is uniquely determined by designating the number of nodes. Fig. 3 is a conceptual view of this model. In the case of an HMM, in particular, a node is generally called a state. In the case shown in Fig. 3, the number of nodes, i.e., the number of states, is four.

The temporary model generating unit 103 determines the number of states of a model in accordance with the number of topics contained in an input text document, and generates a model, i.e., an HMM, in accordance with the number of states. If, for example, it is known that four topics are contained in an input text document, the temporary model generating unit 103 generates only one HMM with four states. If the number of topics contained in an input text document is unknown, the temporary model generating unit 103 generates one each of HMMs with all the numbers of states ranging from an HMM with a sufficiently small number N_{min} of states to an HMM with a sufficiently larger number N_{max} of states (steps 202, 206, and 207). In this case, to generate a model means to ensure a storage area for the storage of the value of a parameter defining a model on a storage medium. A parameter defining a model will be described later.

Assume that the correspondence relationship

between each topic contained in an input text document and each word of the input text document is defined as a latent variable. A latent variable is set for each word. If the number of topics is N, a latent variable can take a value from 1 to N depending on to which topic each word belongs. This latent variable represents the state of a model.

The model parameter initializing unit 104 initializes the values of parameters defining all the models generated by the temporary model generating unit 103 (step 203). Assume that in the case of the above left-to-right type discrete HMM, parameters defining the model are state transition probabilities a_1, a_2,..., a_N and signal output probabilities $b_{1,j}$, $b_{2,j}$,..., $b_{N,j}$. In this case, N represents the number of states. In addition, j = 1, 2,..., L, and L represents the number of types of words contained in an input text document, i.e., the vocabulary size.

A state transition probability a_i is the probability at which a transition occurs from a state i to a state i+1, and $0 < a_i \leqq 1$ must hold. Therefore, the probability at which the state i returns to the state i again is $1-a_i$. A signal output probability $b_{i,j}$ is the probability at which a word designated by an index j is output when the state i is reached after a given state transition. In all states i = 1, 2,..., N, a signal output probability sum total $b_{i,1} + b_{i,2} + \cdots b_{i,L}$

needs to be 1.

The model parameter initializing unit 104 sets, for example, the value of each parameter described above to $a_i = N/T$ and $b_{i,j} = 1/L$ with respect to a model with a state count N. The method to be used to provide this initial value is not specifically limited, and various methods can be used as long as the above probability condition is satisfied. The method described here is merely an example.

The model parameter estimating unit 105 sequentially receives one or a plurality of models initialized by the model parameter initializing unit 104, and estimates a model parameter so as to maximize the probability, i.e., the likelihood, at which the model generates an input text document o_1, o_2,..., o_T (step 204). For this operation, a known maximum likelihood estimation method, an expectation-maximization (EM) method in particular, can be used. As disclosed in, for example, Rabiner et al., (translated by Furui et at.) "Foundation of Sound Recognition (2nd volume)", NTT Advance Technology Corporation, November 1995, pp. 129 - 134 (reference 4), a forward variable $\alpha_t(i)$ and a backward variable $\beta_t(i)$ are calculated throughout t = 1, 2,..., T and i = 1, 2,..., N by using parameter values a_i and $b_{i,j}$ used at this point of time according to recurrent formulas (2). In addition, parameter values are calculated again

according to formulas (3). Formulas (2) and (3) are calculated again by using the parameter values calculated again. This operation is repeated a sufficient number of times until convergence. In this case, δ_{ij} represents a Kronecker delta. That is, if i = j, 1 is set; otherwise, 0 is set.

$$\alpha_1(i) = b_{1,o_1}\delta_{1,i}, \alpha_t(i) = a_{t-1}b_{i,o_t}\alpha_{t-1}(i-1) + (1-a_i)b_{i,o_t}\alpha_{t-1}(i),$$
$$\beta_T(i) = a_N\delta_{N,i}, \beta_t(i) = (1-a_i)b_{i,o_{t+1}}\beta_{t+1}(i) + a_i b_{i+1,o_{t+1}}\beta_{t+1}(i+1) \quad \ldots(2)$$

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1}\alpha_t(i)a_i b_{i+1,o_t}\beta_{t+1}(i+1)}{\sum_{t=1}^{T-1}\alpha_t(i)(1-a_i)b_{i,o_t}\beta_{t+1}(i) + \sum_{t=1}^{T-1}\alpha_t(i)a_i b_{i+1,o_t}\beta_{t+1}(i+1)}, \quad \ldots(3)$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)\delta_{j,o_t}}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)}$$

Convergence determination of iterative calculation for parameter estimation in the model parameter estimating unit 105 can be performed in accordance with the amount of increase in likelihood. That is, the iterative calculation may be terminated when there is no increase in likelihood by the above iterative calculation. In this case, a likelihood is obtained as $\alpha_1(1)\beta_1(1)$. When the iterative calculation is complete, the model parameter estimating unit 105 stores the model parameters a_i and $b_{i,j}$ and the forward and backward variables $\alpha_t(i)$ and $\beta_t(i)$ in the estimation result storage unit 106 in pair with the state counts of models (HMMs) (step 205).

The model selecting unit 107 receives the parameter estimation result obtained for each state count by the model parameter estimating unit 105 from the estimation result storage unit 106, calculates the likelihood of each model, and selects one model with the highest likelihood (step 208). The likelihood of each model can be calculated on the basis of a known AIC (Akaike's Information Criterion), an MDL (Minimum Description Length) criterion, or the like. Information about an Akaike's information criterion and minimum description length criterion is described in, for example, Te Sun Han et al., "Applied Mathematics II of the Iwanami Lecture, Mathematics of Information and Coding", Iwanami Shoten, December 1994, pp. 249 - 275 (reference 5). For example, according to an AIC, a model exhibiting the largest difference between a logarithmic likelihood $\log(\alpha_1(1)\beta_1(1))$ after parameter estimation convergence and a model parameter count NL is selected. In addition, according to an MDL, a selected model is a model whose sum of $-\log(\alpha_1(1)\beta_1(1))$ obtained by sign-reversing a logarithmic likelihood and a product NL x log(T)/2 of a model parameter count and the square root of the word sequence length of an input text document becomes approximately minimum. In the case of both an AIC and an MDL, in general, a selected model is intentionally adjusted by multiplying a term associated with the model parameter count NL by an empirically

determined constant coefficient. It suffices to also perform such operation in this embodiment.

The text segmentation result output unit 108 receives a model parameter estimation result corresponding to a model with the state count N which is selected by the model selecting unit 107 from the estimation result storage unit 106, and calculates a segmentation result for each topic for the input text document in the estimation result (step 209).

By using the model with the state count N, the input text document $o_1, o_2, \ldots, o_T$ is segmented into N sections. The segmentation result is probabilistically calculated first according to equation (4). Equation (4) indicates the probability at which a word o_t in the input text document is assigned to the ith topic section. The final segmentation result is obtained by obtaining i with which $P(z_t = i | o_1, o_2, \ldots, o_T)$ is maximized throughout t = 1, 2, ..., T.

$$P(z_t = i \mid o_1, o_2, \ldots, o_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \quad \ldots(4)$$

In this case, the model parameter estimating unit 105 sequentially updates the parameters by using the maximum likelihood estimation method, i.e., formulas (3). However, MAP (Maximum A Posteriori) estimation can also be used instead of the maximum likelihood estimation method. Information about maximum a

posteriori estimation is described in, for example, Rabiner et al., (translated by Furui et at.) "Foundation of Sound Recognition (2nd volume)", NTT Advance Technology Corporation, November 1995, pp. 166 - 169 (reference 6). In the case of maximum a posteriori estimation, if, for example, conjugate prior distributions are used as the prior distributions of model parameters, the prior distribution of a_i is expressed as beta distribution $\log p(a_i|\kappa_0, \kappa_1) = (\kappa_0 - 1) \times \log(1 - a_i) + (\kappa_1 - 1) \times \log(a_i) + \text{const}$, and the distribution of b_{ij} is expressed as direct distribution $\log p(b_{i,1}, b_{i,2}, \ldots, b_{i,L}|\lambda_1, \lambda_2, \ldots, \lambda_L) = (\lambda_1 - 1) \times \log(b_{i,1}) + (\lambda_2 - 1) \times \log(b_{i,2}) + \cdots + (\lambda_L - 1) \times \log(b_{i,L}) + \text{const}$, where κ_0, κ_1, λ_1, $\lambda_2, \ldots, \lambda_L$ and const are constants. At this time, parameter updating formulas for maximum a posteriori estimation corresponding to formulas (3) for maximum likelihood estimation are expressed as:

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1) + \kappa_1 - 1}{\sum_{t=1}^{T-1} \alpha_t(i)(1-a_i) b_{i,o_t} \beta_{t+1}(i) \kappa_0 - 1 + \sum_{t=1}^{T-1} \alpha_t(i) a_i b_{i+1,o_t} \beta_{t+1}(i+1) + \kappa_1 - 1},$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i) \delta_{j,o_t} + \lambda_j - 1}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i) + \sum_{k=1}^{L} (\lambda_k - 1)} \quad \ldots (5)$$

In this embodiment described so far, the signal output probability b_{ij} is made to correspond to a

state. That is, the embodiment uses a model in which a word is generated from each state (node) of an HMM. However, the embodiment can use a model in which a word is generated from a state transition (arm). A model in which a word is generated from a state transition is useful for a case wherein, for example, an input text is an OCR result on a paper document or a speech recognition result on a speech signal. This is because, in the case of a text document containing a speech pause in a speech signal or a word indicating a line feed in a paper document, i.e., <pause> or <line feed>, if a signal output probability is fixed such that a word generated from a state transition from the state i to the state i+1 is always <pause> or <line feed>, <pause> or <line feed> can always be made to correspond to a topic boundary detected from the input text document by this embodiment. Assume that the input text document is not an OCR result or speech recognition result. Even in this case, if a signal output probability is set in advance such that a word closely associated with a topic change such as "then", "next", "well", or the like is generated from a state transition from the state i to the state i+1 in a model in which a word is generated from a state transition, a word like "then", "next", or "well" can be made to easily appear at a detected topic boundary.

Second Embodiment

The second embodiment of the present invention will be described in detail next with reference to the accompanying drawings.

This embodiment is shown in the block diagram of Fig. 1 like the first embodiment. That is, this embodiment comprises a text input unit 101 which inputs a text document, a text storage unit 102 which stores the input text document, a temporary model generating unit 103 which generates one or a plurality of models each describing the transition between topics of the text document and in which information indicating which word of the text document belongs to which topic is made to correspond to a latent variable and each word of the text document is made to correspond to an observable variable, a model parameter initializing unit 104 which initializes the value of each model parameter which defines each model generated by the temporary model generating unit 103, a model parameter estimating unit 105 which estimates the model parameter of the model initialized by the model parameter initializing unit 104 by using the model and the text document stored in the text storage unit 102, an estimation result storage unit 106 which stores the parameter estimation result obtained by the model parameter estimating unit 105, a model selecting unit 107 which selects a parameter estimation result on one model from parameter estimation

results on a plurality of models if they are stored in the estimation result storage unit 106, and a text segmentation result output unit 108 which segments the input text document in accordance with the parameter estimation result on the model selected by the model selecting unit 107 and outputs the segmentation result. Each unit can be implemented by being operated by a program stored in a computer or by reading the program recorded on a recording medium.

The operation of this embodiment will be sequentially described next.

The text input unit 101, text storage unit 102, and temporary model generating unit 103 respectively perform the same operations as those of the text input unit 101, text storage unit 102, and temporary model generating unit 103 of the first embodiment described above. As in the first embodiment, the text storage unit 102 stores an input text document as a string of words, a string of concatenations of two or three adjacent words, or a general string of concatenations of n words, and an input text document which is written in Japanese having no spaces between words can be handled as a word string by applying a known morphological analysis method to the document.

The model parameter initializing unit 104 initializes the values of parameters defining all the models generated by the temporary model generating unit

103. Assume that each model is a left-to-right type discrete HMM as in the first embodiment, and is further defined as a tied-mixture HMM. That is, a signal output from a state i is linear combination $c_{i,1}b_{1,j} + c_{i,2}b_{2,j} + \cdots c_{i,M}b_{M,j}$ of M signal output probabilities $b_{1,j}$, $b_{2,j}$,..., $b_{M,j}$, and the value of $b_{i,j}$ is common to all states. In general, M represents an arbitrary natural number smaller than a state count N. Information about a tied-mixture HMM is described in, for example, Rabiner et al., (translated by Furui et at.) "Foundation of Sound Recognition (2nd volume)", NTT Advance Technology Corporation, November 1995, pp. 280 - 281 (reference 7). The model parameters of a tied-mixture HMM include a state transition probability a_i, a signal output probability $b_{j,k}$ common to all states, and a weighting coefficient $c_{i,j}$ for the signal output probability. In this case, i = 1, 2,..., N, where N is a state count, j = 1, 2,..., M, where M is the number of types of topics, and k = 1, 2,..., L, where L is the number of types of words, i.e., the vocabulary size, contained in an input text document. The state transition probability a_i is the probability at which a transition occurs from a state i to a state i+1 as in the first embodiment. The signal output probability $b_{i,j}$ is the probability at which a word designated by an index k is output in a topic j. The weighting coefficient $c_{i,j}$ is the probability at which the topic j occurs in the state i.

As in the first embodiment, the sum total $b_{j,1} + b_{j,2} + \cdots + b_{j,L}$ of signal output probabilities needs to be 1, and sum total $c_{i,1} + c_{i,2} + \cdots c_{i,L}$ of weighting coefficients needs to be 1.

The model parameter initializing unit 104 sets, for example, the value of each parameter described above to $a_i = N/T$, $b_{j,k} = 1/L$, and $c_{i,j} = 1/M$ with respect to a model with a state count N. The method to be used to provide this initial value is not specifically limited, and various methods can be used as long as the above probability condition is satisfied. The method described here is merely an example.

The model parameter estimating unit 105 sequentially receives one or a plurality of models initialized by the model parameter initializing unit 104, and estimates a model parameter so as to maximize the probability, i.e., the likelihood, at which the model generates an input text document $o_1, o_2, \ldots, o_T$. For this operation, an expectation-maximization (EM) method can be used as in the first embodiment. A forward variable $\alpha_t(i)$ and a backward variable $\beta_t(i)$ are calculated throughout t = 1, 2,..., T and i = 1, 2,..., N by using parameter values a_i, $b_{j,k}$, and $c_{i,j}$ used at this point of time according to recurrent formulas (6). In addition, parameter values are calculated again according to formulas (7). Formulas (6) and (7) are calculated again by using the parameter values

calculated again. This operation is repeated a sufficient number of times until convergence. In this case, δ_{ij} represents a Kronecker delta. That is, if i = j, 1 is set; otherwise, 0 is set.

$$\alpha_1(i) = \sum_{j=1}^{M} c_{1,j} b_{j,o_1} \delta_{1,j}, \alpha_t(i) = \sum_{j=1}^{M} \{a_{i-1} c_{i,j} b_{j,o_t} \alpha_{t-1}(i-1) + (1-a_i) c_{i,j} b_{j,o_t} \alpha_{t-1}(i)\},$$

$$\beta_1(i) = a_N \delta_{N,j}, \beta_t(i) = \sum_{j=1}^{M} \{(1-a_i) c_{i,j} b_{j,o_{t+1}} \beta_{t+1}(i) + a_i c_{i+1,j} b_{j,o_{t+1}} \beta_{t+1}(i+1)\}$$

...(6)

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1} \sum_{j=1}^{M} \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1)}{\sum_{t=1}^{T-1} \sum_{j=1}^{M} \{\alpha_t(i)(1-a_i) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1)},$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^{T} \sum_{i=1}^{N} \{\alpha_t(i)(1-a_i) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1)\}}{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{L} \{\alpha_t(i')(1-a_{i'}) c_{i',j} b_{j,k} \beta_{t+1}(i') + \alpha_t(i') a_{i'} c_{i'+1,j} b_{j,k} \beta_{t+1}(i'+1)\}},$$

$$c_{ij} \leftarrow \frac{\sum_{t=1}^{T} \{\alpha_t(i)(1-a_i) c_{i,j} b_{j,o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j} b_{j,o_t} \beta_{t+1}(i+1)\}}{\sum_{j=1}^{M} \sum_{t=1}^{T} \{\alpha_t(i)(1-a_i) c_{i,j'} b_{j',o_t} \beta_{t+1}(i) + \alpha_t(i) a_i c_{i+1,j'} b_{j',o_t} \beta_{t+1}(i+1)\}}$$

...(7)

Convergence determination of iterative calculation for parameter estimation in the model parameter estimating unit 105 can be performed in accordance with the amount of increase in likelihood. That is, the iterative calculation may be terminated when there is no increase in likelihood by the above iterative calculation. In this case, a likelihood is obtained as $\alpha_1(1)\beta_1(1)$. When the iterative calculation is complete, the model parameter estimating unit 105

stores the model parameters a_i, $b_{j,k}$, and $c_{i,j}$ and the forward and backward variables $\alpha_t(i)$ and $\beta_t(i)$ in the estimation result storage unit 106 in pair with the state counts of models (HMMs).

The model selecting unit 107 receives the parameter estimation result obtained for each state count by the model parameter estimating unit 105 from the estimation result storage unit 106, calculates the likelihood of each model, and selects one model with the highest likelihood. The likelihood of each model can be calculated on the basis of a known AIC (Akaike's Information Criterion), MDL (Minimum Description Length) criterion, or the like.

In the case of both an AIC and an MDL, as in the first embodiment, a selected model is intentionally adjusted by multiplying a term associated with the model parameter count NL by an empirically determined constant coefficient.

Like the text segmentation result output unit 108 in the first embodiment, the text segmentation result output unit 108 receives a model parameter estimation result corresponding to a model with the state count N which is selected by the model selecting unit 107 from the estimation result storage unit 106, and calculates a segmentation result for each topic for the input text document in the estimation result. A final segmentation result can be obtained by obtaining

i, throughout t = 1, 2,..., T, with which $P(z_t = i|o_1, o_2, \ldots, o_T)$ is maximized, according to equation (4).

Note that, as in the first embodiment, the model parameter estimating unit 105 may estimate model parameters by using the MAP (Maximum A Posteriori) estimation method instead of the maximum likelihood estimation method.

Third Embodiment

The third embodiment of the present invention will be described next with reference to the accompanying drawings.

This embodiment is shown in the block diagram of Fig. 1 like the first and second embodiments. That is, this embodiment comprises a text input unit 101 which inputs a text document, a text storage unit 102 which stores the input text document, a temporary model generating unit 103 which generates one or a plurality of models each describing the transition between topics of the text document and in which information indicating which word of the text document belongs to which topic is made to correspond to a latent variable and each word of the text document is made to correspond to an observable variable, a model parameter initializing unit 104 which initializes the value of each model parameter which defines each model generated by the temporary model generating unit 103, a model parameter estimating unit 105 which estimates the model parameter of the

model initialized by the model parameter initializing unit 104 by using the model and the text document stored in the text storage unit 102, an estimation result storage unit 106 which stores the parameter estimation result obtained by the model parameter estimating unit 105, a model selecting unit 107 which selects a parameter estimation result on one model from parameter estimation results on a plurality of models if they are stored in the estimation result storage unit 106, and a text segmentation result output unit 108 which segments the input text document in accordance with the parameter estimation result on the model selected by the model selecting unit 107 and outputs the segmentation result. Each unit can be implemented by being operated by a program stored in a computer or by reading the program recorded on a recording medium.

The operation of this embodiment will be sequentially described next.

The text input unit 101, text storage unit 102, and temporary model generating unit 103 respectively perform the same operations as those of the text input unit 101, text storage unit 102, and temporary model generating unit 103 of the first and second embodiments described above. As in the same manner in the first and second embodiments of the present invention, the text storage unit 102 stores an input text document as a string of words, a string of

concatenations of two or three adjacent words, or a general string of concatenations of n words, and an input text document which is written in Japanese having no spaces between words can be handled as a word string by applying a known morphological analysis method to the document.

The model parameter initializing unit 104 hypothesizes kinds of distributions by using model parameters, i.e., a state transition probability a_i and a signal output probability b_{ij} as probability variables with respect to one or a plurality of models generated by the temporary model generating unit 103, and initializes the values of the parameters defining the distributions. Parameters which define the distributions of model parameters will be referred to as hyper-parameters with respect to original parameters. That is, the model parameter initializing unit 104 initializes hyper-parameters. In this embodiment, as the distributions of state transition probabilities a_i and signal output probabilities b_{ij}, the following are used respectively: beta distribution $\log p(a_i | \kappa_{0,i}, \kappa_{1,i}) = (\kappa_{0,i}-1) \times \log(1 - a_i) + (\kappa_{1,i}-1) \times \log(a_i) + \text{const}$ and direct distribution $\log p(b_{i,1}, b_{i,2}, \ldots, b_{i,L} | \lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,L}) = (\lambda_{i,1} -1) \times \log(b_{i,1}) + (\lambda_{i,2} -1) \times \log(b_{i,2}) + \cdots + (\lambda_{i,L} -1) \times \log(b_{i,L}) + \text{const}$. The hyper-parameters are $\kappa_{0,1}$, $\kappa_{1,i}$, and $\lambda_{i,j}$. In this case, $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, L$. The model

parameter initializing unit 104 initializes hyper-parameters, for example, according to $\kappa_{0,i} = \kappa_0$, $\kappa_{1,i} = \kappa_1$, and $\lambda_{ij} = \lambda_0$ for $\kappa_0 = \varepsilon(1 - N/T) + 1$, $\kappa_1 = \varepsilon N/T + 1$, and $\lambda_0 = \varepsilon/L + 1$. A proper positive number like 0.01 is assigned to ε. Note that the method to be used to provide this initial value is not specifically limited, and various methods can be used. This initialization method is merely an example.

The model parameter estimating unit 105 sequentially receives one or a plurality of models initialized by the model parameter initializing unit 104, and estimates hyper-parameters so as to maximize the probability, i.e., the likelihood, at which the model generates the input text document o_1, o_2,..., o_T. For this operation, a known variational Bayes method derived from the Bayes estimation method can be used. For example, as described in Ueda, "Bayes Learning [III] - Foundation of Variational Bayes Learning", THE TRANSACTIONS OF THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS, July 2002, Vol 85, No. 7, pp. 504 - 509 (reference 8), a forward variable $\alpha_t(i)$ and a backward variable $\beta_t(i)$ are calculated throughout t = 1, 2,..., T and i = 1, 2,..., N by using hyper-parameter values $\kappa_{0,i}$, $\kappa_{1,i}$, and $\lambda_{i,j}$ obtained at this point of time, and hyper-parameter values are further calculated again according to formula (9). Formulas (8) and (9) are calculated again by using

the parameter values calculated again. This operation is repeated a sufficient number of times until convergence. In this case, δ_{ij} represents a Kronecker delta. That is, if i = j, 1 is set; otherwise, 0 is set. In addition, $\Psi(x) = d(\log \Gamma(x))/dx$, and $\Gamma(x)$ is a gamma function.

$$\begin{aligned}
&\alpha_1(i) = \exp(B_{i,o_1})\delta_{1,i},\\
&\alpha_t(i) = \alpha_{t-1}(i-1)\exp(A_{1,i-1} + B_{i,o_t}) + \alpha_{t-1}(i)\exp(A_{0,i} + B_{i,o_t}),\\
&\beta_T(i) = \exp(A_{1,N})\delta_{N,i},\\
&\beta_t(i) = \beta_{t+1}(i)\exp(A_{0,i} + B_{i,o_{t+1}}) + \beta_{t+1}(i+1)\exp(A_{1,i} + B_{i+1,o_{t+1}})
\end{aligned} \qquad \ldots(8)$$

for

$$\begin{aligned}
&A_{0,i} = \Psi(\kappa_{0,i}) - \Psi(\kappa_{0,i} + \kappa_{1,i}),\\
&A_{1,i} = \Psi(\kappa_{1,i}) - \Psi(\kappa_{0,i} + \kappa_{1,i}),\\
&B_{ik} = \Psi(\lambda_{ik}) - \Psi\left(\sum_{j=1}^{L}\lambda_{ij}\right)
\end{aligned}$$

$$\kappa_{0,i} \leftarrow \kappa_0 + \sum_{t=1}^{T-1}\overline{z_{t,i}z_{t+1,i}},\ \kappa_{1,i} \leftarrow \kappa_1 + \sum_{t=1}^{T-1}\overline{z_{t,i}z_{t+1,i+1}} + \delta_{N,i},\lambda_{ik} \leftarrow \lambda_0 + \sum_{t=1}^{T-1}\overline{z_{t,i}\delta_{k,o_t}} \qquad \ldots(9)$$

for

$$\begin{aligned}
&\overline{z_{t,i}} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)},\\
&\overline{z_{t,i}z_{t+1,i}} = \frac{\alpha_t(i)\exp(A_{0,i} + B_{i,o_{t+1}})\beta_{t+1}(i)}{\sum_{j=1}^{N}\sum_{s=\{0,1\}}\alpha_t(j)\exp(A_{s,j} + B_{j+s,o_{t+1}})\beta_{t+1}(j+s)},\\
&\overline{z_{t,i}z_{t+1,i+1}} = \frac{\alpha_t(i)\exp(A_{1,i} + B_{i+1,o_{t+1}})\beta_{t+1}(i+1)}{\sum_{j=1}^{N}\sum_{s=\{0,1\}}\alpha_t(j)\exp(A_{s,j} + B_{j+s,o_{t+1}})\beta_{t+1}(j+s)}
\end{aligned}$$

Convergence determination of iterative calculation for parameter estimation in the model parameter estimating unit 105 can be performed in

accordance with the amount of increase in approximate likelihood. That is, the iterative calculation may be terminated when there is no increase in approximate likelihood by the above iterative calculation. In this case, an approximate likelihood is obtained as product $\alpha_1(1)\beta_1(1)$ of forward and backward variables. When the iterative calculation is complete, the model parameter estimating unit 105 stores the hyper-parameters $\kappa_{0,i}$, $\kappa_{1,i}$, and $\lambda_{i,j}$ and the forward and backward variables $\alpha_t(i)$ and $\beta_t(i)$ in the estimation result storage unit 106 in pair with the state counts N of models (HMMs).

Note that as a Bayes estimation method in the model parameter estimating unit 105, an arbitrary method such as a known Markov chain Monte Carlo method or Laplace approximation method other than the above variational Bayes method can be used. This embodiment is not limited to the variational Bayes method.

The model selecting unit 107 receives the parameter estimation result obtained for each state count by the model parameter estimating unit 105 from the estimation result storage unit 106, calculates the likelihood of each model, and selects one model with the highest likelihood. As the likelihood of each model, a known Bayesian criterion (Bayes posteriori probability) can be used within the frame of the above variational Bayes method. A Bayesian criterion can be calculated by formula (10). In formula (10), P(N) is the priori

probability of a state count, i.e., a topic count N, which is determined in advance by some kind of method. If there is no specific reason, P(N) may be a constant value. In contrast, if it is known in advance that a specific state count is likely to occur or not likely to occur, P(N) corresponding to the specific state count is set to a large or small value. In addition, as the hyper-parameters $\kappa_{0,i}$, $\kappa_{1,i}$, and $\lambda_{i,j}$ and the forward and backward variables $\alpha_t(i)$ and $\beta_t(i)$, values corresponding to the state count N are acquired from the estimation result storage unit 106 and used.

$$P(N)\alpha_1(1)\beta_1(1)$$
$$\times \exp\left\{\sum_{i=1}^{N}(\kappa_{0,i}-\kappa_0)(\Psi(\kappa_{0,i}+\kappa_{1,i})-\Psi(\kappa_{0,i}))+\sum_{i=1}^{N}(\kappa_{1,i}-\kappa_1)(\Psi(\kappa_{0,i}+\kappa_{1,i})-\Psi(\kappa_{1,i}))\right\}$$
$$\times \exp\left\{\sum_{i=1}^{N}\sum_{k=1}^{L}(\lambda_{ij}-\lambda_0)\left(\Psi\left(\sum_{j=1}^{L}\lambda_{ij}\right)-\Psi(\lambda_{ik})\right)\right\}$$
$$\times \prod_{i=1}^{N}\left\{\frac{\Gamma(\kappa_0+\kappa_1)\Gamma(\kappa_{0,i})\Gamma(\kappa_{1,i})}{\Gamma(\kappa_{0,i}+\kappa_{1,i})\Gamma(\kappa_0)\Gamma(\kappa_1)}\frac{\Gamma\left(\sum_{j=1}^{L}\lambda_0\right)}{\Gamma\left(\sum_{j=1}^{L}\lambda_{i,j}\right)}\prod_{j=1}^{L}\frac{\Gamma(\lambda_{ij})}{\Gamma(\lambda_0)}\right\} \quad \ldots(10)$$

Like the text segmentation result output unit 108 in the first and second embodiments described above, the text segmentation result output unit 108 receives a model parameter estimation result corresponding to a model with the state count, i.e., the topic count N, which is selected by the model selecting unit 107 from the estimation result storage unit 106, and calculates a segmentation result for each topic for the input text

document in the estimation result. A final segmentation result can be obtained by obtaining i, throughout t = 1, 2,..., T, with which $P(z_t = i|o_1, o_2,..., o_T)$ is maximized, according to equation (4).

Note that in this embodiment, as in the second embodiment described above, the temporary model generating unit 103, model parameter initializing unit 104, and model parameter estimating unit 105 can be each configured to generate a tied-mixture left-to-right type HMM, instead of a general left-to-right type HMM, initialize, and perform parameter estimation.

Fourth Embodiment

The fourth embodiment of the present invention will be described in detail next with reference to the accompanying drawings.

Referring to Fig. 4, the fourth embodiment of the present invention comprises a recording medium 601 on which a text-processing program 605 is recorded. The recording medium 601 may be a CD-ROM, magnetic disk, semiconductor memory, or the like, and the embodiment also includes the distribution of the text-processing program through a network. The text-processing program 605 is loaded from the recording medium 601 into a data processing device (computer) 602, and controls the operation of the data processing device 602.

In this embodiment, under the control of the text-processing program 605, the data processing device

602 executes the same processing as that executed by the text input unit 101, temporary model generating unit 103, model parameter initializing unit 104, model parameter estimating unit 105, model selecting unit 107, and text segmentation result output unit 108 in the first, second, or third embodiment, and outputs a segmentation result for each topic with respect to an input text document by referring to a text recording medium 603 and a model parameter estimation result recording medium 604 each of which contains information equivalent to that in a corresponding one of the text storage unit 102 and the estimation result storage unit 106 in the first, second, or third embodiment.